留出法

"留出法" (hold-out)直接将数据集D 划分为两个互斥的集合?其中一个集合作为训练集S，另一个作为测试集T， 即， .在S上训练出模型后，用T来评估其测试误差，作为对泛化误差的估计.

交叉验证法

"交叉验证法" (cross validation)先将数据集D划分为k个大小相似的互斥子集，即 每个子集都

尽可能保持数据分布的一致性，即从D中通过分层采样得到. 然后，每次用k-1个子集的并集作为训练集，余下的那个子集作为测试集;这样就可获得k组训练/测试集，从而可进行k次训练和测试，最终返回的是这k个测试结果的均值。显然，交叉验证法评估结果的稳定性和保真性在很大程度上取决于k的取值。

**自助法**

"自助法" (bootstrapping)是一个比较好的解决方案，它直接以自助采样法(bootstrap sampling) 为基础，给定包含m 个样

本的数据集D ， 我们对它进行采样产生数据集D': 每次随机从D 中挑选一个样本，将其拷贝放入D'，然后再将该样本放回初始数据集D中，使得该样本在下次采样时仍有可能被采到;这个过程重复执行m次后，我们就得到了包含m个样本的数据集D'，这就是自助采样的结果.显然， D 中有一部分样本会在D'中多次出现，而另一部分样本不出现.可以做一个简单的估计，样本在m 次采样中始终不被采到的概率是， 取极限得到 .

**错误率和精度**

错误率是分类错误的样本数占样本总数的比例，精度则是分类正确的样本数占样本总数的比例.

**知识点：准确率，查准率，查全率，真正例率，假正例率**

问题：准确率，查准率，查全率，真正例率，假正例率的定义

**知识点：泪淆矩阵、查准率、查全率与Fl**

**泪淆矩阵**：

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划

分为真正例(true positive) 、假正例(false positive) 、真反倒(true negative) 、

假反例(false negative) 四种情形，令TP 、FP 、TN 、FN 分别表示其对应的

样例数，则显然有TP+FP+TN+FN=样例总数.

查准率和查全率：

查准率P 与查全率R 分别定义为：

查准率和查全率是一对矛盾的度量.一般来说，查准率高时，查全率往往偏低;而查全率高时，查准率往往偏低.

**P-R 曲线**：

根据学习器的预测结果对样例进行排序，排在前面的是学习器认为"最可能"是正例的样本?排在最后的则是学习器认为"最

不可能"是正例的样本.按此顺序逐个把样本作为正例进行预测，则每次可以计算出当前的查全率、查准率以查准率为纵轴、查全率为横轴作图，就得到了查准率-查全率曲线，简称" P-R 曲线"

Fl

Fl 是基于查准率与查全率的调和平均(harmonic mean)定义的:

**ROC与AUC**

ROC 全称是"受试者工作特征" (Receiver Operating Characteristic) 曲线，根据学习器的预测结果对样例进行排序，按此顺序逐个把样本作为正例进行预测，每次计算出两个重要量的值，分别以它们为横、纵坐标作图'就得到了"ROC 曲线”。纵轴是"真正例率" (True Positive Rate，简称TPR) ，横轴是"假正例率" (False PositiveRate，简称FPR) ，两者分别定义为：

AUC (Area UnderROC Curve), 定义为ROC 曲线下各部分的面积之和

**线性模型**

给定由d 个属性描述的示例æ = (X1; 白;...; Xd) ， 其中均是a 在第4 个属

性上的取值，线性模型(linear model)试图学得一个通过属性的线性组合来进行

预测的函数，即

一般用向量形式写成:

W和b学得之后，模型就得以确定.

**最小二乘法线性回归**

线性回归试图学得：

试图让均方误差最小化，即:

基于均方误差最小化来进行模型求解的方法称为"最小二乘法" (least squ町e method). 在线性回归中，最小A乘法就是试图找到一条直线，使所有样本到直线上的欧氏距离之和最小.

**对数几率回归**：

单调可做函数将分类任务的真实标记υ 与线性回归模型的预测值联系起来.对数几率函数(logistic function) 正是这样一个常用的替代函数:

上式中y是概率，对应的对数几率是：

用线性回归模型的预测结果去逼近真实标记的对数几率，因此，其对应的模型称为"对数几率回归" (logistic regression，亦称logit regr臼sio丑) .特别需注意到，虽然它的名字是"回归"，但实际却是一种分类学习方法.这种方法有很多优点，例如它是直接对分类可能性进行建模，无需事先假设数据分布?这样就避免了假设分布不准确所带来的问题;它不是仅预测出"类别"，而是可得到近似概率预测，这对许多需利用概率辅助决策的任务很有用;

LDA:

线性判别分析(Linear Discriminant Analysis，简称LDA) 是一种经典的线性学习方法，，思想非常朴素: 给定训练样例集7 设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近、异类样例的投影点尽可能远离;在对新样本进行分类时，将其投影到同样的这条直线上，再根据投影点的位置来确定新样本的类别.

LDA也常被视为一种经典的监督降维技术

**纠错输出码**

Error CorrectingOutput Codes，简称ECOC。是将编码的思想引入类别拆分，并尽可能在解码过程中具有容错性. ECOC 工作过程主要分为两步:

·编码:对N 个类别做M 次划分， 每次划分将一部分类别划为正类，一部分划为反类，从而形成一个二分类训练集;这样一共产生M 个训练集，可训练出M 个分类器.

·解码:M 个分类器分别对测试样本进行预测，这些预测标记组成一个编码.将这个预测编码与每个类别各自的编码进行比较，返回其中距离最小的类别作为最终预测结果.

**决策树**：

一棵决策树包含一个根结点、若干个内部结点和若干个叶结点; 叶结点对应于决策结果，其他每个结点则对应于一个属性测试;每个结点包含的样本集合根据属性测试的结果被划分到子结点中;根结点包含样本全集.从根结点到每个叶结点的路径对应了一个判定测试序列.决策树学习的目的是为了产生一棵泛化能力强，即处理未见示例能力强的决策树，其基本流程遵循简单且直观的"分而治之" (divide-and-conquer) 策略，

**信息熵**

"信息铺" (information entropy)是度量样本集合纯度最常用的一种指标.

假定当前样本集合D 中第k 类样本所占的比例为Pk *(k* = 1, 2,. . . , IYI) ，则D

的信息娟定义为：

Ent(D) 的值越小，则D 的纯度越高.

**信息增益**

假定离散属性α 布V 个可能的取值，"信息增益" (information gain)定义为：

信息增益越大，则意味着使周属性α 来进行划分所获得的"纯

度提升"越大.因此，我们可用信息增益来进行决策树的划分属性选择

**增益率**

**基尼指数**

**机器学习**

假设用P 来评估计算机程序在某任务类T 上的性能，若一个程序通过利用经验E 在T 中任务丰获得了性能改善，则我们就说关于T 和P ， 该程序对E 进行了学习.

**模型与模式（model & pattern）**

. "模型"泛指从数据中学得的结果有文献用"模型"指全局性结果(例如一棵决策树)，而用"模式"指局部性结呆(例如A条规则).

**独立同分布**

通常假设样本空间中全体样本服从一个未知"分布" D，获得的每个样本都是独立地从这个分布上采样获得。

**泛化能力**

学得模型适用于新样本的能力，称为"泛化" (generalization) 能力.具有强泛化能力的模型能

很好地适用于整个样本空间.

**假设空间和版本空间**

我们可以把学习过程看作一个在所有假设(hypothesis)组成的空间中进行搜索的过程，搜索目标是找到与训练集"匹配"但t) 的假设，即能够将训练集中的瓜判断正确的假设.假设的表示一旦确定，假设空间及其规模大小就确定了.

可能有多个假设与训练集一致，即存在着一个与训练集一致的"假设集合"，我们称之为"版本空间" (version space).

**归纳偏好**

机器学习算法在学习过程中对某种类型假设的偏好，称为"归纳偏好" (inductive bias) ,

或简称为"偏好"。任何一个有效的机器学习算法必有其归纳偏好，否则它将被假设空间中看

似在训练集上"等效"的假设所迷惑，而无法产生确定的学习结果. 归纳偏好可看作学习算法自身在一个可能很庞大的假设空间中对假设进行选择的启发式或"价值观"

**奥卡姆剃刀**

"奥卡姆剃刀" (Occam's razor)是一种常用的、自然科学研究中最基本的原则，即"若有多个假设与观察一致，则选最简单的那个"